Lab #6
Multiple testing

In this lab, we will be working with an Affymetrix data set that was run on the human HGU95A array. This experiment was designed to assess the gene expression events in the frontal cortex due to aging. A total of 18 male and 12 female postmortem brain samples were obtained to assess this.

The analysis that we are interested in conducting is a direct follow up to the previous lab of differential expression. We first want to identify those genes/probes that are differentially expressed in the frontal cortex between old and young subjects, then between males and females. Next, we would like to evaluate the differences between a couple of multiple testing adjustment methods. As explained in the lecture and the course website, multiple testing is a necessary step to reduce false positives when conducting more than a single statistical test. You will generate some p-value plots to get an idea of the how conservative some methods are compared to others.

I have identified 2 gene vectors for you to use below, so do not calculate the t-test or adjustments on the entire array of genes/probes.

For the second part of this lab, you will be working with RNA-sequencing data from The Cancer Genome Atlas (TCGA), specifically a breast invasive carcinoma dataset of 119 patient tumors. The data matrix and annotation files are on the course website. We will be trying to confirm an observation from a meta-analysis performed by Mehra et al, 2005 in Cancer Research. The authors identified the gene (using arrays) and protein (using immunohistochemistry) GATA3 as a prognostic factor in breast cancer, where patients with low expression of GATA3 experienced overall worse survival. The PubMed abstract is here: http://www.ncbi.nlm.nih.gov/pubmed/16357129.

1.) Download the GEO Brain Aging study from the class website. Also obtain the annotation file for this data frame.

2.) Load into R, using read.table() function and the header=T/row.names=1 arguments for each data file.

```
> lab6 <- read.table("/Users/stevendea/Desktop/JHU/Fall 2019/Gene Expression Data Analysis and Visualization/Labs/Lab 6/agingStudy11FCortexAffy.txt", header = T, row.names = 1)
```

3.) Prepare 2 separate vectors for comparison. The first is a comparison between male and female patients. The current data frame can be left alone for this, since the males and females are all grouped together. The second vector is comparison between patients >= 50 years of age and those < 50 years of age.
To do this, you must use the annotation file and logical operators to isolate the correct arrays/samples.

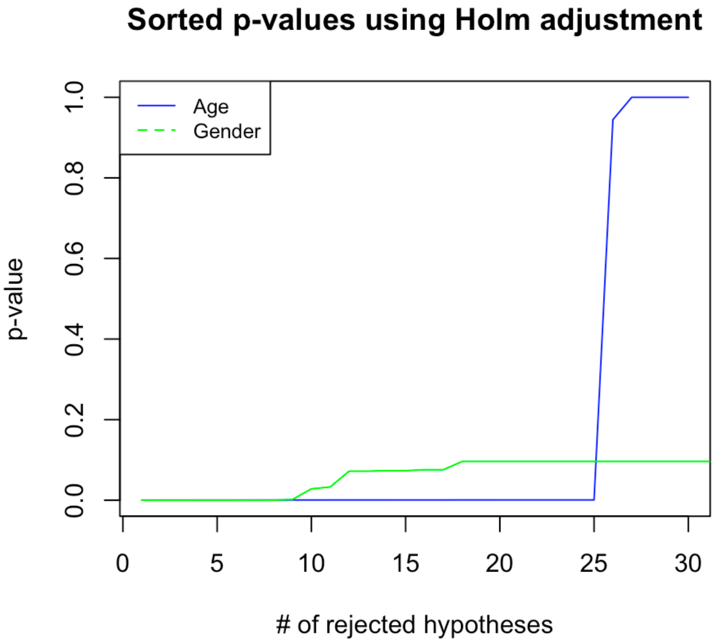
```
> lab6.ann <- read.table("/Users/stevendea/Desktop/JHU/Fall 2019/Gene Expression Data Analysis and Visualization/Labs/Lab 6/agingStudy1FCortexAffyAnn.txt", header=T, row.names = 1)
> g.g
> g.a
> lab6.o50 <- which(lab6.ann$Age > 50)
> lab6.u50 <- which(lab6.ann$Age < 50)
```

4.) Run the t.test function from the notes using the first gene vector below for the gender comparison. Then use the second gene vector below for the age comparison. Using these p-values, use either p.adjust in the base library or mt.rawp2adjp in the multtest library to adjust the values for multiple corrections with the Holm's method.

```
# first gene vector comparison
> lab6.1 <- lab6[g.g,]
> lab6.1.18 <- c(1:18) # male indexes
> lab6.1.30 <- c(19:30) # female indexes
> rawp <- apply(lab6.1,1,t.test.all.genes,s1=lab6.1.18,s2=lab6.1.30)

# second gene vector comparison
> lab6.2 <- lab6[g.a,]
> lab6.o50 <- which(lab6.ann$Age > 50) # older than 50 indexes
> lab6.u50 <- which(lab6.ann$Age < 50) # under 50 indexes
> rawp2 <- apply(lab6.2,1,t.test.all.genes,s1=lab6.o50,s2=lab6.u50)

# adjust both sets of p-values using the holm method
> adjustedp <- p.adjust(rawp, method = "holm")
> adjustedp2 <- p.adjust(rawp2, method = "holm")
```

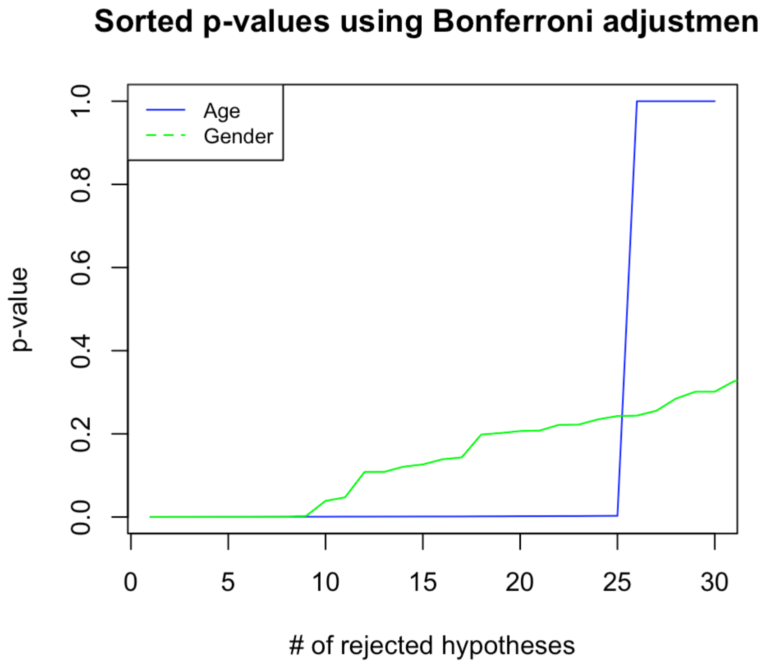
5.) Sort the adjusted p-values and non-adjusted p-values and plot them vs. the x-axis of numbers for each comparison data set. Make sure that the two lines are different colors. Also make sure that the p-values are sorted before plotting.

```
# sort the adjusted p-values
> adjustedp.sorted <- sort(adjustedp)
> adjustedp2.sorted <- sort(adjustedp2)

# plot the p-values vs. # of rejected hypotheses for the Age and Gender t-test
> plot(adjustedp2.sorted, type = "l", col = "blue", ylab = "p-value", xlab = "# of
rejected hypotheses", main = "Sorted p-values using Holm adjustment")
> lines(adjustedp.sorted, col = 'green')
> legend("topleft", legend=c("Age", "Gender"), col = c("blue", "green"), lty=1:2,
cex=0.8)
```



6.) Repeat #4 and #5 with the Bonferroni method.

```
# adjust using Bonferroni method
> adjustedp.bon <- p.adjust(rawp, method = "bonferroni")
> adjustedp2.bon <- p.adjust(rawp2, method = "bonferroni")

# sort adjusted Bonferroni p-values
> adjustedp.bon.sorted <- sort(adjustedp.bon)
> adjustedp2.bon.sorted <- sort(adjustedp2.bon)

# plot
> plot(adjustedp2.bon.sorted, type = "l", col = "blue", ylab = "p-value", xlab = "#
of rejected hypotheses", main = "Sorted p-values using Bonferroni adjustment")
> lines(adjustedp.bon.sorted, col = 'green')
> legend("topleft", legend=c("Age", "Gender"), col = c("blue", "green"), lty=1:2,
cex=0.8)
```



7.) Read in the log$_2$ normalized fragments per kb per million mapped reads (FPKM) data matrix and annotation files. This is RNA-sequencing data that has normalized read counts on a similar scale to microarray intensities.

```
#Read in data and annotation file
> lab6.7 <- read.table("/Users/stevendea/Desktop/JHU/Fall 2019/Gene Expression
Data Analysis and Visualization/Labs/Lab 6/tcga_brca_fpkm.txt", header=T,
row.names = 1)

> lab6.7.ann <-read.table("/Users/stevendea/Desktop/JHU/Fall 2019/Gene
Expression Data Analysis and Visualization/Labs/Lab 6/tcga_brca_fpkm_sam.txt",
header=T, row.names=1, fill=TRUE)
```

8.) Use grep to subset the data matrix only by gene 'GATA3' and make sure to cast this vector to numeric.

```
# search for GATA3 pattern in the row names of the data matrix
> lab6.GATA3 <- as.numeric(grep("GATA3", row.names(lab6.7)))
[1] 6362
```

9.) Create a binary (1/0) vector for the patients where the **upper** 25% expression of GATA3 is coded as 1 and all other patients are coded as 0. Call this new variable 'group'.

```
# sort all of the data of GATA3 in reverse order
> GATA3 <- lab6.7[lab6.7.GATA3,]
> GATA3.sorted <- sort(GATA3, decreasing = TRUE)

# of 119 values, where is the cutoff for the top 25%?
> 119*0.25
[1] 29.75 # round down

# which value is the lowest value of the top 25% of the GATA3 vector
> GATA3.sorted[29]
        TCGA.H4.A2HO.01A.11R.A180.07
GATA3|2625                12.99197

# ifelse to create a new binary vector if the value is in the top 25%
> group = ifelse(GATA3 > 12.99197, yes = 1, no = 0)

# create a new vector of the indexes where true (1)
> group1 <- which(group == 1)
 [1]   1   4   7  11  13  21  24  25  27  38  51  56  58
[14]  59  60  62  64  67  71  77  82  83  85  88  90  91
[27]  97 117 119
```

10.) Create a data matrix with the ‘group’ variable you created in #9 and the remaining variables in the annotation file.

```
# subset lab6.7.ann matrix using the group vector from #9
> group1.matrix <- lab6.7.ann[group1,]

# subset lab6.7.ann matrix using everything NOT in the group vector from #9
> group2 <- which(group == 0)
> group2.matrix <- lab6.7.ann[group2,]
```

11.) Run a Kaplan-Meier (KM) analysis to determine if a difference in survival experience exists between the two GATA3 expression groups using the survdiff function. Extract the p-value from the chi squared test output.

```
# change the vital_status column to numeric and casting all 0’s (alive) to 1’s(dead)
and all 2’s to 1 before running through survdiff()
> group1.matrix$vital_status <- as.numeric(group1.matrix$vital_status)
> group1.vitals <- which(group1.matrix$vital_status == 1)
> group1.matrix$vital_status[group1.vitals] = 1
> group1.vitals2 <- which(group1.matrix$vital_status == 2)
> group1.matrix$vital_status[group1.vitals2] = 0

# run survdiff on group1 (high GATA3 expression)with time = age_at_initial, status
= vital_status, event = months_to_event
> survdiff(Surv(group1.matrix$age_at_initial_pathologic_diagnosis,
group1.matrix$vital_status) ~group1.matrix$months_to_event)

Chisq= 47.5  on 27 degrees of freedom, p= 0.009

# run survdiff on group 2 (lower GATA3 expression)
> group2.matrix$vital_status <- as.numeric(group2.matrix$vital_status)
> group2.vitals <- which(group2.matrix$vital_status == 1)
> group2.matrix$vital_status[group2.vitals] = 1
> group2.vitals2 <- which(group2.matrix$vital_status == 2)
> group2.matrix$vital_status[group2.vitals2] = 0
> survdiff(Surv(group2.matrix$age_at_initial_pathologic_diagnosis,
group2.matrix$vital_status) ~group2.matrix$months_to_event)

Chisq= 243  on 70 degrees of freedom, p= <2e-16
```

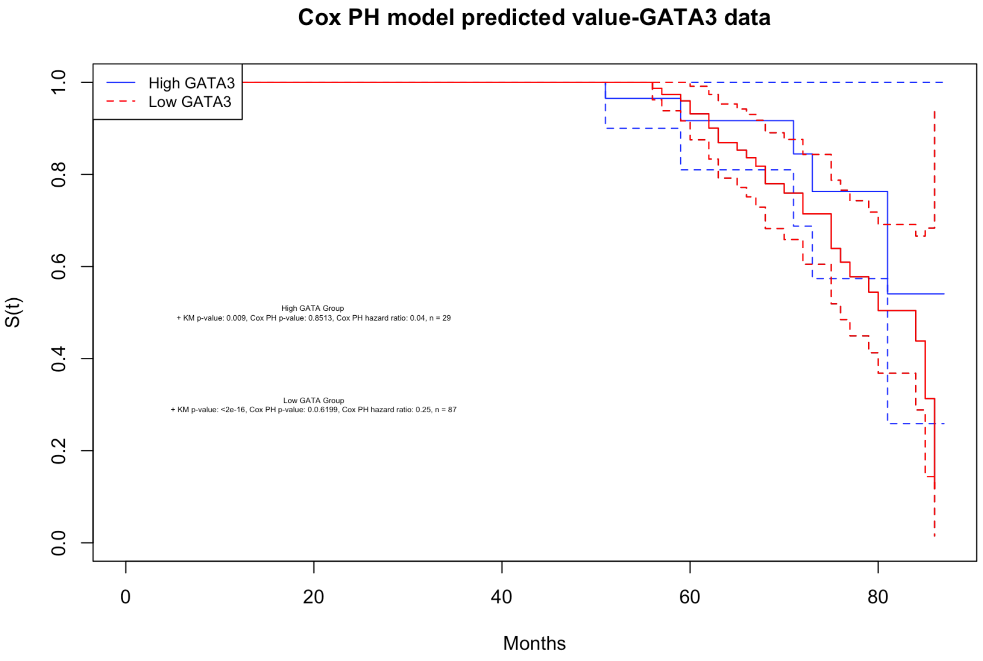
12.) Now run a Cox proportion hazard (PH) regression model on just the grouping variable (i.e. no other covariates) and extract both the p-value and hazard ratio from the output.

```
> fit1 <- coxph(Surv(group1.matrix$age_at_initial_pathologic_diagnosis,
group1.matrix$vital_status) ~group1.matrix$months_to_event)

Likelihood ratio test=0.04  on 1 df, p=0.8513
n= 29, number of events= 5
```

```
> fit2 <- coxph(Surv(group2.matrix$age_at_initial_pathologic_diagnosis,
group2.matrix$vital_status) ~group2.matrix$months_to_event)

Likelihood ratio test=0.25  on 1 df, p=0.6199
n= 87, number of events= 27
```

13.) Run the survfit() function only on the grouping variable (i.e. no other covariates) and plot the KM curves, being sure to label the two groups with a legend, two different colors for each line, and provide the KM p-value, Cox PH p-value, Cox PH hazard ratio, and sample sizes all in each of the two groups all on the plot.

```
> plot( survfit(fit1),xlab="Months",ylab="S(t)",main="Cox PH model predicted
value-GATA3 data", col= 'blue')
> lines(survfit(fit2))
> lines(survfit(fit2), col = 'red')
> legend("bottomleft", legend = c("High GATA3", "Low GATA3"), col = c('blue',
'red'), lty=1:2, cex=0.8)

# Add text to the bottom
> temptext1 <- "High GATA Group
+  KM p-value: 0.009, Cox PH p-value: 0.8513, Cox PH hazard ratio: 0.04, n = 29"
> temptext2 <- "Low GATA Group
+  KM p-value: <2e-16, Cox PH p-value: 0.0.6199, Cox PH hazard ratio: 0.25, n =
87"

> text(20,0.5, temptext1,cex=0.4)
> text(20,0.3, temptext2,cex=0.4)
```



14.) Does this result agree with the Mehra et al, study result?

It is somewhat similar to the Mehra et al study result in that the GATA3 high group had a higher “rate” of survival and survived for longer on average, which ties in with Mehra’s conclusions that GATA3 is highly associated with maintaining differentiation of breast cells, thus keeping them out of a progenitor-like cancer state.

Gene Vectors (indices for specific rows/genes)

```
# gender comparison gene vector
g.g <- c(1394, 1474, 1917, 2099, 2367, 2428, 2625, 3168, 3181, 3641, 3832, 4526,
4731, 4863, 6062, 6356, 6684, 6787, 6900, 7223, 7244, 7299, 8086, 8652,
8959, 9073, 9145, 9389, 10219, 11238, 11669, 11674, 11793)

# age comparison gene vector
g.a <- c(25, 302, 1847, 2324, 246, 2757, 3222, 3675, 4429, 4430, 4912, 5640, 5835,
5856, 6803, 7229, 7833, 8133, 8579, 8822, 8994, 10101, 11433, 12039, 12353,
12404, 12442, 67, 88, 100)
```